Laboratory information system data extraction and re-use: opportunities and challenges

Christopher R. McCudden[1], Matthew P. A. Henderson[2]

[1]Department. of Pathology & Lab. Medicine, Division of Biochemistry, [2]Department. of Pediatrics, Division of Metabolics, University of Ottawa, Ottawa, Canada

Correspondence to: Christopher R. McCudden, PhD, DABCC, FACB, FCACB. Clinical Biochemist, Division of Biochemistry, The Ottawa Hospital, Ottawa, Canada; Associate Professor, Department. of Pathology & Lab. Medicine, University of Ottawa, 501 Smyth Rd. Ottawa, ON K1H 8L6, Canada. Email: cmccudden@toh.on.ca.

Abstract: Laboratory information systems (LISs) are a rich source of data. LIS data can be used for numerous purposes including operations, quality projects, and research. LIS data can inform decision making, provide value additions, and ultimately be used to improve patient care. However, there are many challenges that come with LIS data re-use. These include security, access to information, the ability to analyze large volumes of data, data quality, and validation. Herein we describe the pros and cons of LIS data re-use and provide a framework for a typical LIS data extraction. Also provided are concrete examples where LIS data was essential and beneficial for a successful project. Collectively, laboratorians need to focus on training initiatives to empower future staff who will require these skills to do their jobs effectively. Laboratorians should also ask more of the LIS vendors in terms of data access and analytical tools.

Keywords: Informatics; laboratory information systems (LISs); error detection; quality assurance; laboratory operations

Received: 29 August 2017; Accepted: 19 September 2017; Published: 17 October 2017.
doi: 10.21037/jlpm.2017.09.07
View this article at: http://dx.doi.org/10.21037/jlpm.2017.09.07

Introduction

Each day, laboratories generate thousands of results. This data is rich with analytical, patient demographic, physician order, temporal, and patient location information. Laboratory information system (LIS) data can be re-used for many purposes such as operations, quality, and research. LIS data can guide organizational decisions, help detect errors, improve reference intervals, and facilitate discovery of areas for quality improvement. However, LIS has largely been designed for one-way transactions, in the form to getting information in, rather than getting information out. As a result, there is an array of challenges in using LIS data. Challenges include access, extraction, analysis, and validation where it can be difficult to get, use, and harvest actionable information. This manuscript describes the rich opportunity that laboratory information provides as well as the dark side of acquiring and re-using data from LIS.

Opportunities for LIS data re-use

There are numerous opportunities for LIS data re-use. Consider that each LIS result contains information about the patient, ordering physician, lab results, test order, as well as date, time, and encounter location. We routinely find numerous uses for this information for operations, quality and research.

Operations

From an operational standpoint, LIS data it can be used to make decisions about basic workflow, such as when to send couriers, when to add or reduce laboratory staff, and how fast results get from one location to another. It can also be used for basic business planning around instrument and hardware replacement.

On a daily basis, turnaround time is commonly assessed

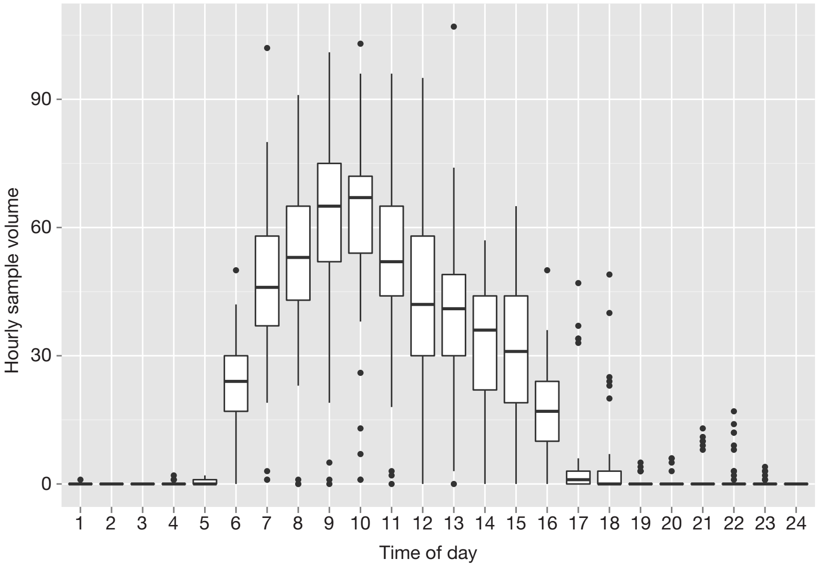


Figure 1 Example of sample processing volume data visualization. Volume represents weekday hematology and biochemistry samples collected at a one hospital site.

with built-in LIS applications. While these might serve a basic need, we have found it very helpful to be able to pull out raw data to calculate more robust statistics, such as median, trimmed means, and standard deviation. In addition, the ability to visualize trends and patterns without software limitations is essential for a deep understanding. With raw data in hand, additional analysis is enabled, where a model can be built to assess turnaround time and determine whether there is a significant difference from previous performance. Where data is available in real time, home-brew algorithms may be used to flag samples to staff to identify and resolve problems before they result in negative outcomes or delays.

As a basis for instrument replacement or changes in technology and methods, it is very useful to identify common test users. We have found it particularly helpful in terms of who to direct communications to when a particular test has a problem. In this way, rather than making assumptions about the source and origin of orders, data will dictate where to target communication ensuring that key individuals are not missed.

As a multi-site facility, we have used available LIS data to determine when best to send couriers between sites. Here, visualization of data can be very informative in terms of when the maximum specimen flow occurs (*Figure 1*). This information serves the basis for when couriers should be sent. LIS data in the form of test volumes can also be very useful for instrument replacements and identifying maximal workload, for example to estimates how many instruments, centrifuges, and preanalytical sampling modules to install on an automation line. LIS data can also be used to identify the frequency of downtimes and clinical needs in order to determine when to perform routine maintenance. In this way, the laboratory has their own detailed information rather than relying on vendors or 3rd parties to come in and identify areas of improvement for them.

Basic LIS data is also useful to identify times and locations where demand on staff is high or low. For example, we've experienced instances where staff complain that there is too much work to perform for the available staff. Review of the workload identified only one to two specimens over the course of several hours, thereby confirming that minimal staffing was reasonable. Conversely, we've also identified instances where the staffing is far too low for the workload, which has resulted in either adding additional staff or trying to shift the workload where results are not needed immediately. This is another area that is amenable to data visualization, which be readily identifies issues with workload.

As a final operational use, LIS data may also be used for billing purposes and for reagent contracts. For example, where cost per reportable test is contracted, volumes need to be tallied. LIS data extraction may eliminate manual volume counts and save substantial time and tedious effort.

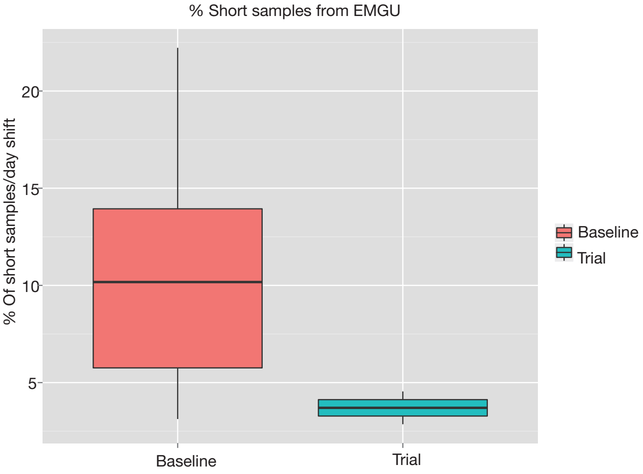


Figure 2 Comparison of the percent of samples with low volumes (short) before and during an experimental trial. During the trial a phlebotomist was stationed in the emergency department.

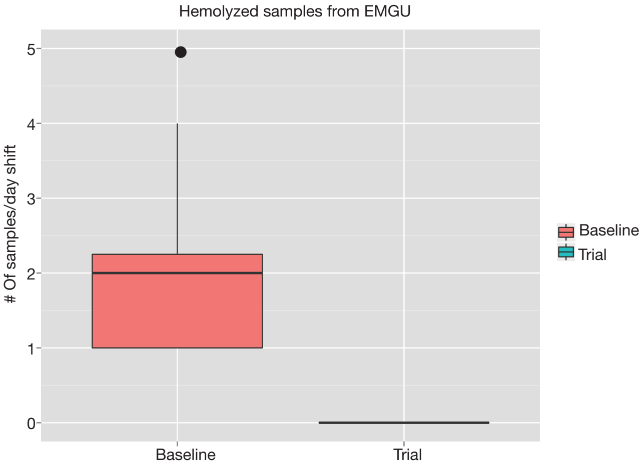


Figure 3 Comparison of number of hemolyzed samples during an experimental trial. During the trial a phlebotomist was stationed in the emergency department.

Quality

Another rich area available for data re-use is quality. For quality purposes we've used LIS data to identify preanalytical errors, to verify and establish reference intervals, and for patient-based quality control.

Two commonly encountered preanalytical errors are hemolysis (rupture of red cells) and short samples (inadequate tube fill volume) due to poor phlebotomy technique. Hemolyzed samples can either yield inaccurate results or prevent reporting of any results, potentially delaying treatment (1). Short samples can result in instrument downtime after clogging of probes due to the gel in PST and SST tubes (plasma/serum separator tubes). We have used LIS data analysis to identify the emergency department (ED) as the primary source of hemolyzed and short samples. Based on this preliminary data, we did a study in collaboration with the ED to test the effect of having a phlebotomist in the ED. Using LIS data, we identified significant improvement in the frequency of short (*Figure 2*) and hemolyzed samples (*Figure 3*). This approach can be used on a prospective basis to target educational efforts and improvement initiatives. In another project, open access to the LIS data facilitated e-mailing of automated reports directly to units to reduce the frequency of preanalytical errors. The utility and flexibility of open LIS data access is essential to these projects.

Another beneficial re-use of LIS data is for confirmation and development of reference intervals. Several studies have identified methods which rely on patient data to generate and confirm the appropriateness of implemented reference intervals (2-4). This is particularly beneficial given the challenges and expenses of drawing samples from healthy volunteers on a routine basis. The availability of LIS data essentially allows continuous monitoring of reference interval.

As a real time quality measure, LIS data can be used to identify shifts and trends in assay performance, classically in the form of moving averages. While many LISs have simple moving averages available, few if any provide enough flexibility to use alternative and more sophisticated methods, such as moving medians, moving variance, and moving deltas (5). Because most LISs have fairly primitive monitoring parameters, extracting data from the system can be helpful for error detection. With complete data, elaborate quality monitoring may be done with multivariate analysis (6). Multivariate analysis may offer better signal to noise for error detection and identify subtler or complex patterns that single moving averages would otherwise miss.

Last but not least, laboratory utilization is greatly facilitated by the availability of LIS data. Beginning with a few basic fields, such as tests, provider, and time, patterns are readily identified that can be used to focus efforts at utilization. Beyond simple descriptive analysis, high level audits can help identify areas of miss-use and direct efforts to apply test controls, limits, feedback, and educational initiatives. Indeed, providing data back to the providers who order tests can be a very effective method for audit and feedback to implement changes and test utilization initiatives.

Research

Beyond operations and quality, there are an infinite number of research projects that rely on LIS data. Some examples of research initiatives we've engaged in include error detection, prediction of urgent dialysis, as well as reference interval generation. While many of these projects focus on basic quality initiatives, there are also opportunities for more elaborate projects such as epidemiological studies. This is particularly true where large healthcare organizations or multicentre facilities have LIS data readily available. Some examples of more advanced research projects we've used LIS data for include establishing autoverification codes, multivariate error detection, and text mining.

In summary, there is no limit to the utility of LIS data. Fundamentally, availability of this information is key to many quality and improvement initiatives as well as answering simple operations and workflow questions. Exploration of information tends to lead to many ideas and solutions, such that there is a real benefit to investing in data access.

Challenges of LIS data re-use

Despite the numerous opportunities and promise of LIS data re-use, there are many challenges. These include access, security, availability, LIS software, data quality, and software/analytical skills.

Access

Access to LIS data is often the first challenge for data reuse. The basis for access problems are manyfold. First, is authorization for access the software system itself, where only select personnel might have the necessary permissions to use the data. This can be common in hospital environments where there is separation between those who administrate and maintain the system and those who need access to the data in it. In reality, healthcare data access comes with important security and privacy concerns. Indeed, providing data access and availability comes with risk. Last year's rash of ransomware attacks (http://www.bbc.com/news/technology-35880610) highlights the additional risk of connected systems, where one ignorant user may cause a system-wide lockdown that could cost the organization a fortune. This high risk scenario frequently leads to use of virtualization environments, such as Citrix, which make system administration easier, but make data extraction harder and often slower. Of course, none of these systems are immune. For many hospitals, software may be outdated, making them particularly susceptible to attack.

Extraction

Beyond security risks and with permissions in hand, the next common hurdle in LIS data re-use is the software. LIS software often presents challenges in terms of ease of extraction of raw information. Consider that LISs are primarily designed for transmission of information into the system rather than extraction of information out. It's only within the last decade or two that the systems have become sources of information to be extracted for secondary analysis. As a result, the design of LIS software is often limited in terms of accessing large volumes of data effectively. Of the systems we've used to date, none have had complete, fast, and easy access. For example, many queries may take too long to be useful because of the design of the system or the hardware on which it's built. The availability of database integration into external software applications is extremely limited if not unknown. Database access from outside of the LIS software is advantageous because it facilitates LIS queries to be part of analytical coding and analysis rather than as part of a patchwork of several separate scripts used to extract, transform, and load information into a database before any queries or analysis can be done.

Analysis

In the rare instance that data is fully available and can be queried by authorized personnel, there remains the limitation of analytical skills. Most laboratory and medical scientific staff are not trained to query databases and analyze large datasets. For example, common spreadsheet software, such as Excel, is not capable of effectively crunching millions or rows of data. This is compounded by the common principle of hiring and promoting people from laboratory positions into LIS positions in the absence of computer, data science, statistical, mathematical, or analytical backgrounds. It will require a new breed of personnel to take advantage of the LIS data re-use opportunity. Analysis of large LIS datasets requires a new set of software tools, analytical skills, and programming capabilities, which are not part of most education and training programs.

Because of the challenges and limitations of skill sets required for extracting data, some LIS vendors have added separate modules for analysis and extraction, often at significant financial expense to the customer. In the last decade there has been the emergence of or entire industry to take advantage of this gap. There is now a host of software and software as service entities and consultants who will readily sell analytical packages, services, and dashboards to willing laboratories and hospitals. In our opinion, training programs have a responsibility to hospitals and laboratories to teach the necessary skills for analysis of LIS datasets. This skill set is the future of the "value addition proposition" on which laboratories are now focused as part of being an information provider rather than a commoditized result provider (7). Further, hospitals and laboratories should demand ready access to their own data from LIS vendors. To this end, when choosing an LIS vendor, laboratorians should carefully consider the hardware and software capabilities needed for data re-use. Patients and taxpayers should be recruited as ready partners in advocating for easier data access, as it will both reduce the need for unnecessary repeat laboratory testing and save money.

Validation

Finally, it cannot be overstated that getting the data is only the beginning. LIS data is often riddled with errors, which range from a simple non-standard datetime format to encountering a completely wrong physician name, which misidentifies who ordered the test. Thus, once data is extracted, the starting point is extensive validation. This is very challenging, as data must be validated from original order (sometimes as a paper requisition) all the way into the electronic record and extracted data itself. Tracing the accuracy of this information can be very time consuming and difficult, requiring defining what the truth should be for detecting data integrity issues.

Identification of problems may still be insufficient, as often data users will encounter system problems that can't be easily addressed. For example, it is common to find a workflow that is the underlying cause of an LIS data integrity issue, which may be insurmountable. Consider trying to get an entire department to change their ordering practices to yield accurate data for re-use that is not to their clear benefit.

Another substantial problem encountered is when LISs are used across a group of hospitals or regional laboratory. Each of these may (will) be built differently, with different order names, test abbreviations, physician naming conventions, and encounter definitions (8). Despite the existence of logical observation identifiers names and codes (LOINC) (9) and HL7, there is limited adoption of standardized terms and naming LIS data. Integrating and exchanging source data from more than one system is extremely difficult and may require use of mapping tables that need to be accurately generated and then maintained prospectively. Much like test harmonization and standardization, this will again be a long road as laboratories recognize this limitation and are only at the beginning of trying to standardize those types of information. Here again, vendors could help in driving standards in LIS builds and naming conventions.

Data extraction background and basics

With a framework of knowledge around the opportunities and challenges of LIS data, this section provides some examples of which fields to extract data from an LIS, and where and how to store data once it's extracted. The goal here is to begin to address the educational needs of those who generate LIS data and may not yet be able to capture of take advantage of it.

Data extraction & structured query language

Most LISs are built on a foundation of a relational database. A relational database refers to a series of tables which are linked together by "keys". A key is a field that can be used to join tables together. In a simple example, the table "test" is joined to the table "container" through the "analyte" field (*Figure 3*). The reason that the information is split into different tables, rather than one wider one, is that with a large amount of data there would be extensive duplication of terms. Duplication takes up a large amount of physical disk space and is slower to query. Envision how many duplications there would be if each creatinine result had the reference interval, units, analytical measuring limits, and test abbreviation recorded in every row.

Relational databases are readily queried using simple commands using a syntax known as SQL. For example, to get all of the information from the tables "test" and "container" the command would be "SELECT * FROM test, container WHERE test.tube=container.tube" (*Table 1*).

Table 1 Query output

Test	Test code	Specimen type	Units	Tube	Color	Volume	Catalog ID
Copper	Cu	Plasma	µmol/L	PST	Mint	5 mL	#F64249X
Cortisol	Cort	Serum	nmol/L	SST	Gold	5 mL	#L22233X
Creatinine	Cr	Serum	µmol/L	SST	Gold	5 mL	#L22233X
Cyclosporine	CSA	EDTA-plasma	µg/L	EDTA	Purple	5 mL	#N64241X

Table 2 Refined query output

Test	Test code	Specimen type	Units	Tube	Color	Volume	Catalog ID
Creatinine	Cr	Serum	µmol/L	SST	Gold	5 mL	#L22233X

In this example, all the tests are joined to all the containers, creating a wider table. Additional "WHERE" commands can be used to refine the query, for example to get all the test and tube data that has the word "creatinine" in it would be as follows: "SELECT * FROM test, tube WHERE test.analyte=tube.analyte AND test='creatinine'" (*Table 2*). In this instance, a single row is yielded by the query.

The underlying structure of all SQL queries relies on the basic principles shown above. However, complexity arises from the number and size of the tables. For example, some LIS, such as Cerner Millennium, have more than 5,000 different tables, making it challenging to find fields and tables of interest. This very complicated table structure is a reflection of an enterprise level hospital information system, of which the LIS is part, and the primary goal is storing information rather than extracting it efficiently. It is here where users must rely on clear and effective documentation. Experience is very useful, and there are often user forums to help laboratorians identify relevant fields and tables and develop effective queries. Indeed, entire courses are dedicated to SQL, but it is not beyond anyone with basic analytical skills, knowledge, and motivation. With some LIS, such as EPIC Beaker, the software provides an abstraction to SQL queries for information. These might provide point and click type interfaces with more user-friendly overviews of useful fields and tables. While these abstractions can facilitate access for casual users, they may inhibit more complex queries that advanced users may need.

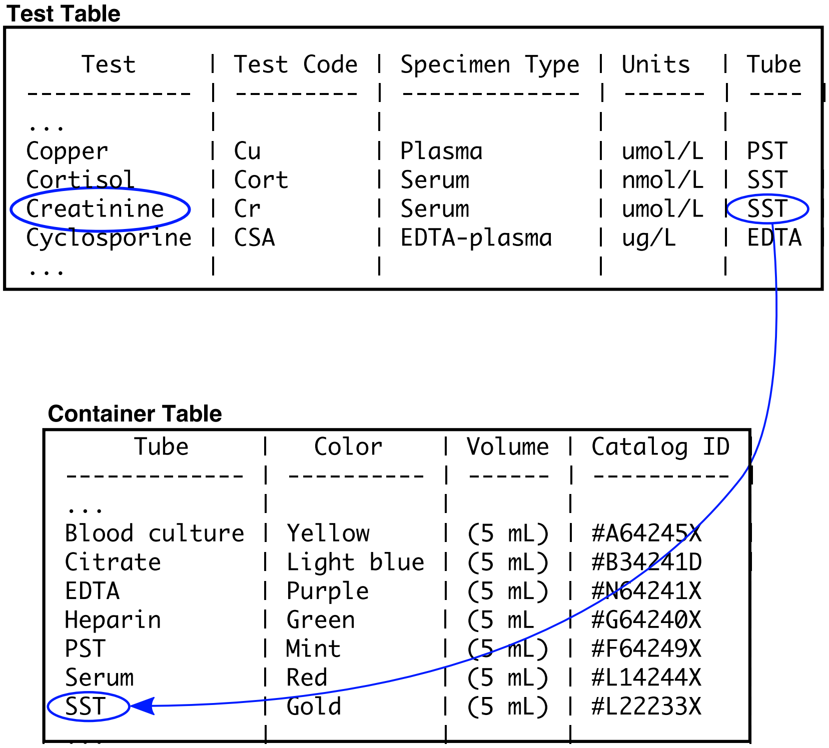
Data transformation and storage

With queries to obtain information of interest in hand (*Figure 4*), the next step is to get the data into the analysis software. This can go in several directions. The first would be to go immediately towards analysis if the LIS software had useful analysis tools available. Most have some basic tools to allows for calculating aggregates and generating summary reports. Where the LIS software tools are insufficient (this is 99% of the time in our experience), the next step would be to either transfer the data directly into a relevant software package (we use the statistical programming language R extensively) or to transfer and upload into a separate database or external datamart.

Lab DataMart

Hospital and LISs have evolved over time to facilitate patient care and laboratory services. One consequence of this evolution is an extremely complex data model that requires extensive training. Happily, the data requirements to support laboratory decision making at the operational and clinical levels are fairly modest. The data elements in *Table 3* capture adequate information for most if not all of the analysis described here.

The fields of *Table 3* can be thought of as a LIS datamart. A datamart is an approach to providing a simplified data extract to end users to facilitate data analysis. In a datamart, only the required data is present in database schema with logical relationships. For example, the core of our datamart would be the sample table. Each sample would be linked to a patient who would be linked to a physician. Each sample would also be linked to multiple analytes. There are a few approaches to creating and using LIS datamarts. Some systems will allow creation of database "views" that could



Test Table

```
   Test      | Test Code | Specimen Type | Units  | Tube
------------ | --------- | ------------- | ------ | ----
...          |           |               |        |
Copper       | Cu        | Plasma        | umol/L | PST
Cortisol     | Cort      | Serum         | nmol/L | SST
Creatinine   | Cr        | Serum         | umol/L | SST
Cyclosporine | CSA       | EDTA-plasma   | ug/L   | EDTA
...          |           |               |        |
```

Container Table

```
     Tube      |   Color    | Volume | Catalog ID
-------------  | ---------- | ------ | ----------
...            |            |        |
Blood culture  | Yellow     | (5 mL) | #A64245X
Citrate        | Light blue | (5 mL) | #B34241D
EDTA           | Purple     | (5 mL) | #N64241X
Heparin        | Green      | (5 mL  | #G64240X
PST            | Mint       | (5 mL) | #F64249X
Serum          | Red        | (5 mL) | #L14244X
SST            | Gold       | (5 mL) | #L22233X
...            |            |        |
```

Figure 4 Relational database structure example. The ellipses represent additional rows for other tests and container types.

Table 3 Minimal tables (bold) with affiliated fields (indented) required for a functional LIS data mart

Patient
ID number
Date of birth
Sex
Physician
ID number
Name
Specialty
Analyte
Name
Result
Flags
Verification date-time
Comments
Sample
Location
Encounter
Reference interval
Ordered date-time
Received data-time
Priority
Accession number

act as a datamart. Data bases views are virtual tables that can be queried directly within the LIS. The another more labour intensive approach is to create a separate database housed outside the LIS. We've used this latter approach extensively to facilitate fast and uninhibited access to the data.

Creating a separate database to house your LIS datamart has some profound advantages in terms of flexibility, and continuity. A stand alone LIS datamart permits querying the database directly from third party reporting and data analysis software. This in turn facilitate interactive analysis and permits automated report generation. In addition, external data sources, such as QC and transfusion medicine results, can be incorporated into the datamart schema. Finally, the datamart provides a level of abstraction between the reporting and analysis infrastructure and the underlying LIS. In the event of a change to the LIS, it is easier to validate the data transmitted from the LIS to the datamart than it is to validate every report based on the datamart.

Summary

LIS data is a veritable gold mine of information. It can improve quality, inform operations, and serve as a foundation for translational research. However, much like gold mining, extracting information can be challenging and potentially dangerous. Laboratories and hospitals who pay millions of dollars for their LIS should pay careful attention to availability of their own data and invest wisely in those who will need the skillsets to access and analyze data.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editors (Tony Badrick and Tze Ping Loh) for the series "Clinical Database in Laboratory Medicine Research Column" published in *Journal of Laboratory and Precision Medicine*. The article has undergone external peer review.

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jlpm.2017.09.07). The series "Clinical Database in Laboratory Medicine Research Column" was commissioned by the editorial office without any funding or sponsorship. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Heyer NJ, Derzon JH, Winges L, et al. Effectiveness of practices to reduce blood sample hemolysis in EDs: a laboratory medicine best practices systematic review and meta-analysis. Clin Biochem 2012;45:1012-32.
2. Hoffmann RG. Statistics in the Practice of Medicine. JAMA 1963;185:864-73.
3. Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. Biometrics 1967;23:115-35
4. Bolann BJ. Easy verification of clinical chemistry reference intervals. Clin Chem Lab Med 2013;51:e279-81.
5. Cervinski M, Cembrowski G. Detection of Systematic Error Using the Average of Deltas. Am J Clin Pathol 2017;147:S165.
6. Leen TK, Erdogmus D, Kazmierczak S. Statistical error detection for clinical laboratory tests. Conf Proc IEEE Eng Med Biol Soc 2012;2012:2720-3.
7. Price CP, John AS, Christenson R, et al. Leveraging the real value of laboratory medicine with the value proposition. Clin Chim Acta 2016;462:183-6.
8. Overhage JM, Evans L, Marchibroda J. Communities' Readiness for Health Information Exchange: The National Landscape in 2004. J Am Med Inform Assoc 2005;12:107-12.
9. Forrey AW, McDonald CJ, DeMoor G, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem 1996;42:81-90.

doi: 10.21037/jlpm.2017.09.07
Cite this article as: McCudden CR, Henderson MP. Laboratory information system data extraction and re-use: opportunities and challenges. J Lab Precis Med 2017;2:81.